

11028
.M414
no. 206-
66

WORKING PAPER

ALFRED P. SLOAN SCHOOL OF MANAGEMENT

William M. ^{artin}Evan and James R. Miller III
Sloan School of Management
Massachusetts Institute of Technology

206-66

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139



Preliminary Draft
October, 1965

DIFFERENTIAL EFFECTS ON RESPONSE BIAS OF COMPUTER VS. CONVENTIONAL
ADMINISTRATION OF A SOCIAL SCIENCE QUESTIONNAIRE:
A LABORATORY EXPERIMENT

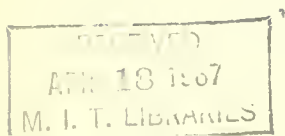
William M. ^{artin}Evan and James R. Miller, III
Sloan School of Management
Massachusetts Institute of Technology

206-66

H320

.M414

13.20.28



DIFFERENTIAL EFFECTS ON RESPONSE BIAS OF COMPUTER VS. CONVENTIONAL
ADMINISTRATION OF A SOCIAL SCIENCE QUESTIONNAIRE:
A LABORATORY EXPERIMENT

Response bias is a subtle phenomenon which is difficult to detect and even more difficult to control. It consists of a tendency on the part of respondents, whether in the role of an interviewee, a testee, or an experimental subject in social science research situations, to give responses which involve either conscious or unconscious distortion of data. The more common term, "response set," refers to specific types of biasing behavior and, hence, is a special case of response bias.

Social scientists have encountered and have sought to control this phenomenon in applying diverse methodologies. Psychologists have struggled with response bias in constructing and administering psychological tests. Sociologists engaged in survey research have also attended to this phenomenon, as have social psychologists who conduct laboratory experiments. It has also received dramatic attention in an appendix, "How to Cheat on Personality Tests," in William Whyte's Organization Man.¹ Whatever the form of response bias, we run the risk of gathering data whose validity is impaired unless such bias can be detected and controlled.

The purpose of this paper is to report the findings of a laboratory experiment on the differential biasing effects of filling out a social

science questionnaire by conventional, pencil and paper procedures versus a new method of typing answers directly onto a teletype console attached to a real-time computer.

DIMENSIONS OF RESPONSE BIAS

In two pioneering articles, Cronbach identified various "response sets." By response sets, Cronbach means "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form."² Among the response sets he identified are the tendency to gamble, i.e., to respond when doubtful; the acquiescence set or the tendency to respond "true" or "agree" to true-and-false questions and agree-or-disagree questions; and the speed versus accuracy set, namely, an inclination to respond rapidly versus carefully in order to achieve a high score. He summarized extensive evidence demonstrating that such response sets as bias in favor of a particular alternative, tendency to guess, and working for speed rather than accuracy operate in many psychological tests.³ Another response set, social desirability, or the tendency to answer in a manner to conform to social norms or to give a favorable impression of oneself to others, has been studied at length by Edwards.⁴

Social scientists concerned with survey research methodology have observed the effect of the auspices under which a questionnaire is administered. If the questionnaire is administered by a particular member of an organization, e.g., by a supervisor or by some other re-

presentative of management, responses are likely to be more favorable about the organization than if the same questionnaire is administered by a non-partisan or by a non-member of the organization, such as a professor. In other words, the social situation surrounding the administration of a questionnaire may be conducive to bias on the part of respondents. If the respondent is doubtful about the confidentiality of the data he is asked to provide, or if he is doubtful about the use to which the information will be put, he may be inclined to bias the results in order to protect his interests in the social situation.⁵ Obviously, if the tendency to bias responses is conscious, as in the case of Whyte's advice to cheat on personality tests, or even if it is unconscious, in either case, the validity of the data is reduced.

Social psychologists conducting laboratory experiments have encountered the tendency of experimental subjects to cooperate or help the experimenter confirm his hypotheses. This disposition to do the experimenter's bidding, as Orne discovered, introduces a biasing effect which can distort the results of an experiment.⁶ Sociologists and anthropologists engaged in field research have observed a similar tendency on the part of respondents. As Vidich and Bensman put it:

"The respondent attempts to form an image of the interviewer and the organization he represents and to form a basis of response with respect to the interviewer.... As a result of this image, respondents provide information which will enable the researcher to solve the particular problem."⁷

Analogous to the findings of Orne and Vidich and Bensman is a study by Jones of the "courtesy bias" encountered in surveys in

Southeast Asia.⁸ For a variety of cultural reasons, social scientists have encountered the courtesy bias, or the wish to please, in conducting interviews and in administering questionnaires. Jones considers various questions concerning the design of research instruments that counteract the courtesy bias. For example, projective drawings have been used to overcome the disinclination to express negative evaluations. Also, phrases designed to establish the acceptability of impolite responses have been used as introductory remarks. An example would be, "Many people around here tell me that...." In addition to concern with questionnaire design, Jones indicates the importance of reassuring anonymity.

"Preventing the undesirable influences of courtesy upon survey responses involves strict attention to other aspects of survey work beyond questionnaire design. The standard reassurances of anonymity must be enforced throughout the interview, since it can be assumed that the respondent's willingness to be 'impolite' will be in direct proportion to the belief that this impoliteness will never be reported in relation to him personally. Interviewers are instructed to interrupt respondents who may wish to give their names; and in some instances of the use of administered questionnaires to special audience groups, the phrase, 'You are not allowed to sign your name' has been inscribed at the top of the questionnaire."⁹

Cronbach and others have recommended a variety of procedures to combat response sets. For example, the use of a multiple-choice test instrument helps to overcome the acquiescence set observed in true-and-false or agree-and-disagree questions. As Cronbach puts it, "response sets are reduced by any procedure that increases the structuration of the test situation."¹⁰

Another method of structuring the test situation which thus far has not been investigated by social scientists as a means of reducing response bias is the use of a computer in administering a psychological test or a survey questionnaire. If such biases as carelessness, lack of candor, and disposition to respond in a socially desirable way for fear of negative evaluation and/or other forms of reprisal operate in psychological tests and survey questionnaires, in part because the interview or the research situation is not confidential or is not perceived as confidential and anonymous, then any procedure that conveys a feeling of privacy could very well reduce response bias. It was this consideration that led us to design our experiment. Our over-all purpose in conducting this experiment was to determine whether or not the administration of questionnaires by means of a computer would elicit more honest and more candid responses than when the same questions were answered in the conventional, pencil and paper manner. Our major hypotheses were as follows:

- 1) Whenever the content of a question is regarded by the subject as highly personal and possibly disturbing, he will respond with greater honesty and candor under computer administration as opposed to conventional administration of the questionnaire. This hypothesis is based on the assumption that subjects will regard typing answers directly into a computer as a situation guaranteeing them greater privacy, greater anonymity, greater confidentiality, and greater freedom from negative evaluations than the conventional situation wherein test papers must be scrutinized,

scored, and interpreted directly by other human beings.

2) When an emotionally neutral and impersonal question is asked, no significant differences in response tendencies attributable to mode of questionnaire administration will appear.

PROCEDURE

Questionnaire Construction

Our first task was to select an appropriate set of questions to ask. We purposely set out to construct a questionnaire whose contents would be regarded by subjects as highly personal and which would succeed in arousing some anxiety and reluctance on their part to respond openly and honestly. Only in this manner could we expect to detect whatever differences in response tendencies might exist between the two modes of administration. In addition, we wished to draw as much as possible from the body of existing psychological test instruments so as to benefit from the experience of previous researchers in questionnaire construction. This led us to select the following items for inclusion in our questionnaire:

- 1) The entire Allport-Vernon-Lindzey (AVL) test of individual values (45 multiple-rank items);
- 2) Questions from the Minnesota Multiphasic Personality Inventory (MMPI) relating to manifest anxiety (20 true-false items);
- 3) Questions from the MMPI Lie Scale designed to detect dishonest responses (15 true-false items);

4) Questions constituting the Srole Scale of perceived socio-cultural anomie (5 multiple-choice items).¹¹

Fifteen true-false questions were then formulated by the authors to determine an individual's felt need for privacy (see Appendix I). The purpose of including these questions was to find out whether differences in response bias tendencies might be correlated with different levels of the need for privacy.

In addition, it was decided to introduce a degree of experimental control into the questionnaire itself, i.e., certain neutral questions were formulated which were relatively impersonal in their content and which were not expected to arouse much anxiety or reluctance to answer on the part of subjects. It was presumed that non systematic differences between modes of administering the questionnaire would be detected on these neutral question items. If differences appeared in the previously listed items where we expected such differences, but failed to appear on these neutral items where we expected no differences, this outcome would lend substantial credibility to our original hypotheses. Neutral items included in our questionnaire were:

- 1) Factual questions constructed by the authors which were considered relatively easy to answer (10 true-false items, see Appendix II).
- 2) Questions constructed by the authors to elicit relatively impersonal opinions (10 true-false items, see Appendix III).

The entire questionnaire contained 120 items--45 items from the AVL test, 35 items from the MMPI, 35 items constructed by the authors, and 5 items constituting the Srole Scale. The 45 items from the AVL test were presented first in their standard order, then the 35 MMPI items and the 35 items constructed by the authors were presented in a scrambled order, and the questionnaire ended with the 5 Srole items.¹²

Sample Selection and Experimental Design

Sample selection was guided by two over-all considerations. First, we wished to collect as homogeneous a sample as possible in order to minimize extraneous variation in differences between modes of administering the questionnaire. Second, for ease of administration, we wished to draw our sample from an available population. We, therefore, decided to recruit M.I.T. undergraduates and to limit the experiment to seniors whose fathers were business or professional men and whose own career objectives were similarly directed.

Notices were sent to each of the undergraduate fraternities requesting volunteers for the experiment. Volunteer forms were distributed to the various fraternities to gather background information on the subjects and to aid in scheduling the experiment. On the basis of volunteer forms returned, a total sample of 60 subjects was chosen, all of whom satisfied the above requirements. The next task was to split this total sample into two sub-samples--one to be given the experimental treatment or computer administration of our questionnaire, the other to be given the control treatment or standard, pencil and paper ad-

ministration. However, this raised a methodological problem.

It was feared at the outset of the experiment that, even if our hypothesis about the bias-reducing effect of computer administration were confirmed, the real significance of such a finding would be mitigated as the general public becomes increasingly more knowledgeable of and experienced with computers. Increased knowledge and experience would probably reduce credibility in the computer's claim to preserve anonymity and confidentiality. Consequently, we considered a two-factor experiment with computer knowledge and experience as the second factor. However, in order to assign subjects in advance to the experimental and control treatments, respectively, and yet still insure a reasonable balance between these two groups in terms of computer knowledge and experience, it would have been necessary to ascertain the degree of each subject's familiarity with computers prior to the experiment. This, we feared, might bias subjects' responses by indicating to them what the experiment was really about. Therefore, we abandoned the idea of a two-factor experiment and relied upon a modified random assignment procedure to balance the two treatment samples with respect to computer knowledge and experience.

The modified random assignment procedure was carried out as follows. First, each subject's major field of concentration was ascertained from his volunteer form. Then, subjects were divided into two groups according to whether this major was electrical engineering or something else. This separation was made on the assumption that electrical engineering majors would probably possess a relatively high degree of computer know-

ledge and experience as compared to others in the sample. Electrical engineering majors were then divided randomly, but evenly, between the two treatments, and the rest of the subjects were assigned at random. This procedure generated an experimental and a control group, each containing 30 subjects.

Questionnaire Administration

Administration of the questionnaire was carried out in three steps. First, both experimental and control subjects were given pre-experimental instructions. For the control subjects, this was a simple statement that they would be asked to fill out a questionnaire with pencil and paper and that their answers would be treated anonymously and confidentially. Without further ado, control subjects were then directed to a private office to complete their task.

For the experimental subjects, a somewhat more elaborate procedure was required. They were told that they would be asked to answer some questions, and they were assured of the same anonymity and confidentiality as were the control subjects. However, experimental subjects were told that their answers were not to be written down on the questionnaire itself. Instead, their answers were to be typed directly into a computer by means of an on-line teletype console. It was further stressed that the printed record of each experimental subject's responses to the various questions could be removed from the teletype console at the end of the experiment and taken home by the subject. This statement was

designed to give experimental subjects additional assurance of anonymity and confidentiality. Next, instructions were given to each subject concerning the proper use of the teletype console. A short practice program was written in FORTRAN to give experimental subjects experience with the console, to test their understanding of the instructions, and to point out any errors in usage.¹³

The second step was carried out by the subjects themselves. Control subjects were left alone in a private office to complete their questionnaires at their own speed. They were requested to inform the experimenter when they were finished and to give him their written results at that time. Experimental subjects were also left alone in a private office with a teletype console. They were requested to type answers directly into the console as instructed by the computer program and to inform the experimenter when they were finished.

A third step consisted of a post-experimental questionnaire administered uniformly to both experimental and control subjects. The post-experimental questionnaire (see Appendix IV) requested every subject to review each of the preceding questions and to indicate:

- 1) Whether or not he had found that question disturbing in any way;

- 2) Whether or not he had been reluctant to give a candid answer.

In addition, questions were asked to ascertain:

- 1) How persuaded each subject was that his answers would in fact be anonymous and confidential;

- 2) How serious each subject had been in expressing his opinions;

- 3) How many semester-length courses each subject had taken which

were directly concerned with digital computer technology;

4) How extensive each subject would rate his knowledge of computers;

5) How extensive each subject would rate his first-hand experience with computers.

After completing the post-experimental questionnaire, subjects were given an explanation of the rationale of the experiment, thanked for their cooperation, and paid for their efforts.

RESULTS

Specific Predictions

Experimental measures relevant to testing our primary hypotheses were of two types. First, there were personal questions included within the questionnaire for the express purpose of inducing anxiety and reluctance on the part of all subjects to answer honestly. We anticipated that these questions would in fact be perceived by everyone as personal and at least somewhat disturbing, but we anticipated that experimental subjects, feeling assured of greater privacy, anonymity, and confidentiality, would answer these questions with greater honesty and candor than would control subjects.

Neutral questions included within the questionnaire constituted a second set of relevant measures. Few subjects were expected to perceive these items as either personal or disturbing, nor was much reluctance to respond with candor anticipated. Consequently, very little

response bias was anticipated, and, therefore, no differences in response bias between experimental and control subjects were predicted on these neutral items.

Listed below for easy comparison with actual results are the specific predictions made concerning these two types of measures prior to our experiment.

1) Experimental subjects would have a higher average score than control subjects on the religious value orientation or the AVL test. In the social environment of a college undergraduate (particularly at an engineering school), religious values and sentiments are probably attenuated. Consequently, the more honest and candid response would show up as a greater willingness to admit a positive orientation toward religious values. No predictions were made concerning AVL results on other than the religious value orientation.

2) Similarly, experimental subjects would admit symptoms of manifest anxiety more freely than would control subjects. This would show up as a higher proportion of favorable responses for experimental subjects on the Manifest Anxiety Scale of the MMPI.

3) Experimental subjects would register a smaller proportion of dishonest responses than control subjects on the Lie Scale of the MMPI. This constitutes our most direct test of the primary hypothesis regarding differential response bias.

- 4) Experimental subjects would agree more readily than control subjects to socially undesirable statements indicating socio-cultural anomie. This would manifest itself as a higher agreement score on the Srole Scale.
- 5) There would be no significant differences between experimental and control subjects on either the factual questions or the impersonal opinion questions purposefully introduced into the questionnaire as a neutral comparison base.

Test of Predictions

Our findings on the religious value score of the AVL test agreed with our first prediction. Experimental subjects achieved a mean score of 34.87, while control subjects registered a mean score of 27.50. The difference of 7.37 between the two group means resulted in a t-value of 3.51, which is statistically significant at the .0005 level (one-tail).

Results relating to our other four predictions are summarized in Table 1. The statistics reported are the proportion of responses

Insert Table 1 about here.

favorable to the indicated scale given by the experimental group (first column), the proportion of responses favorable to the same scale given by the control group (second column), the difference in proportions (third column), and the statistical significance of the indicated difference (last column).¹⁴

Inspection of Table 1 shows that all but one of the predictions were confirmed. The experimental subjects admitted more symptoms of manifest anxiety, told fewer lies, and agreed more readily with the Srole statements of anomie--all as predicted. Non-systematic differences appeared between experimental and control subjects on the Factual and Impersonal Opinion Scales--also as predicted. However, a relatively large difference did appear on the Impersonal Opinion Scale which turned out to be significant at the .011 level. This was definitely not predicted, nor could it be interpreted in the framework of our original hypotheses. We shall revert to this result later when we discuss the impact of the acquiescence set.

A more sensitive test was designed to distinguish the cases where differences were predicted from the cases where no differences were predicted. The Lie Scale was first reversed (i.e., to indicate number of truths told rather than number of lies told), and then pooled with the Manifest Anxiety and Srole Scales. A weighted average proportion of favorable responses on this pooled sample was computed both for experimental and for control subjects. Similarly, the Factual and Impersonal Opinion Scales were pooled, and a weighted average proportion of favorable responses was computed both for experimental and for control subjects. These results appear in Table 2.

Insert Table 2 about here.

Inspection of Table 2 affords clear confirmation of our hypothesized results. Where directional predictions were made (i.e., on the Manifest Anxiety, Lie, and Srole Scales), a small but highly significant weighted average difference emerged between the two proportions, and this difference was in the predicted direction. In contrast, where no differences were predicted (i.e., on the Factual and Impersonal Opinion Scales), a much smaller and insignificant difference emerged.¹⁵

Although highly encouraging, the above results constitute only a partial confirmation of our hypotheses. Merely to show that predicted results actually occurred is not enough. It must also be shown that the predicted results occurred for the reasons stated in the original hypotheses--not just by accident or for some other reasons. In the case of our hypotheses, this means demonstrating that:

- 1) Questions contained in the religious portion of the AVL test, the Manifest Anxiety Scale, the Lie Scale, and the Srole Scale really were regarded by the entire group of subjects as more personal and disturbing than questions contained in the rest of the AVL test, the Factual Scale, and the Impersonal Opinion Scale.
- 2) Experimental subjects really were more thoroughly persuaded that their responses would be treated as anonymous and confidential than were control subjects.

Responses to some of the post-experimental questions were used to test whether or not each of the above two conditions did in fact hold during

the experimental situation.

The reader will recall that all subjects were asked to review the entire questionnaire after completion of the experiment and to indicate which of the component questions had disturbed them and which ones they were reluctant to answer honestly. As anticipated, the religious portion of the AVL test, the Manifest Anxiety Scale, the Lie Scale, and the Srole Scale were regarded as far more disturbing than the rest of the AVL test, the Factual Scale, and the Impersonal Opinion Scale. Whereas the former set of personal questions succeeded in disturbing subjects 584 out of 2200 instances, or 26.6 percent of the time, the latter set of neutral questions were regarded as disturbing only 343 out of 2400 instances, or 14.3 percent of the time. This difference was significant at well beyond the .001 level and served to demonstrate the first condition listed above.¹⁶

Regarding the second condition, experimental subjects appeared to have been slightly more persuaded than controls that their answers would be treated anonymously and confidentially. On the persuasion question in the post-experimental questionnaire, where individual scores could range from one (meaning very much persuaded) to seven (meaning not at all persuaded), the mean score registered by experimental subjects was 2.655, while the mean score registered by control subjects was 2.933. The difference between these two means was in the predicted direction. However, it was neither large nor very significant. A t-test indicated a t value of .69, significant at the

.247 level (one-tail).

One further measure of the validity of subjects' responses was provided by the seriousness question in the post-experimental questionnaire. On a scale similar to the one associated with the persuasion question, the entire group of subjects registered an average score of 1.609, suggesting that our subjects were quite serious in expressing their opinions. This result appeared to hold equally well for both experimental and control subjects. Their group means differed by only .225 points.

Having achieved reasonable confirmation of our original hypotheses, we then proceeded to test for a second-order effect. This involved the privacy questions included in the experimental questionnaire. We reasoned that, if a feeling of greater privacy, anonymity, and confidentiality induces greater honesty and candor, then perhaps the extent to which this effect occurs might be directly related to an individual's felt need for privacy. That is, the greater an individual's felt need for privacy, the greater might be the effect of giving him additional privacy upon inducing more honest and candid responses.

To test this idea, a rank-order correlation was performed between each individual's total score on the Privacy Scale and his total scores on the Manifest Anxiety Scale, the Lie Scale, and the Srole Scale, respectively. The results were not confirmatory. Kendall tau correlation coefficients ranged in magnitude from $-.169$ to $+.130$,

and associated significance levels ranged from .115 to .580. In other words, an essentially random pattern of correlations emerged. Hence, we set aside the notion of this second-order effect.

Impact of the Acquiescence Set

Let us now return to the anomalous significant difference on the Impersonal Opinion Scale reported in Table 1. As stated previously, no significant difference was predicted on this scale; yet one occurred. This raised the possibility of some joint effect brought about by the conditions of the experimental treatment in conjunction with the acquiescence set. Specifically, we reasoned as follows. A response was scored as favorable to the Impersonal Opinion Scale if the subject agreed with the stated opinion and indicated this by checking "true." Otherwise, if he checked "false," that response was scored as unfavorable to the scale. Consequently, it would be impossible to separate an acquiescence response (i.e., blind acceptance without regard to the content of the opinion) from a genuine agreement response on any of the impersonal opinion questions.

If computer administration of the questionnaire should happen to induce or intensify the acquiescence set, then a significantly larger number of agreements by experimental subjects than by control subjects on the Impersonal Opinion Scale might very well indicate a differential tendency to acquiesce rather than genuine differences in opinion between the two groups. If true, this would serve to confuse our other findings, since the Lie Scale, the Srole Scale, and all but five items

on the Manifest Anxiety Scale were similarly scored, i.e., checking "true" was uniformly scored as favorable to the scale. For this reason, it became quite important to investigate the possible connection between acquiescence and the computer administration of the questionnaire.

The first step in isolating the possible effects of the acquiescence set was to find a way of distinguishing a blind acceptance response from a genuine agreement. This could easily be done on the Factual Scale, since, in constructing that scale, factually correct answers had been randomized with respect to the true-false categories. Therefore, a direct test of the acquiescence phenomenon and, more particularly, of the connection between acquiescence and the computer administration of the questionnaire could be performed as follows.

- 1) Focus attention upon responses to questions contained in the Factual Scale.
- 2) Identify those items in the Factual Scale where the factually correct response is to disagree with the statement or to check it "false."
- 3) Cross-classify responses to these items first according to whether they were given by an experimental or by a control subject, and then according to whether they were checked "true" or "false."
- 4) If there were any differential tendency for computer administration to induce acquiescence, this should manifest itself as a greater frequency on the part of experimental subjects to check

"true," even though this is a factually incorrect response.

5) If no such connection exists between acquiescence and computer administration, then experimental and control subjects should check "true" with approximately the same frequency.

Table 3 contains a cross-tabulation of the above information.

Insert Table 3 about here.

Inspection of Table 3 indicates that experimental subjects checked "true" slightly more frequently than did control subjects, even though this was the wrong answer. However, a Chi Square analysis performed on Table 3 resulted in a Chi Square value of .379, which was significant at only the .269 level (one-tail). A binomial test performed on the top-row frequencies of Table 3 yielded an almost identical one-tail significance level of .279. These results did not give adequate support to the existence of a tendency for computer administration to induce acquiescence.

One additional test was performed using five items contained in the Manifest Anxiety Scale. Whereas all of the Lie Scale and Srole Scale items, as well as fifteen out of the twenty Manifest Anxiety items, were constructed so that a "true" response was scored as favorable to the scale, five of the Manifest Anxiety items were scored in reverse, viz., a response checked as "false" was scored as favorable to the scale. This provided the basis for an additional check on the possible connection between computer administration of the questionnaire and acquiescence.

If the results reported in Table 2 were really the result of acquiescence, induced by computer administration, then the weighted average difference of .041 between experimental and control subjects should at least disappear and perhaps even reverse sign when attention is focused exclusively upon the five reverse-order Manifest Anxiety items. But actual analysis of these five items showed a weighted average difference of .033. The difference and its sign were preserved, although at a slightly reduced magnitude. The significance of this difference was .246 (one-tail).

In light of the above three results, it did not seem reasonable to attribute the anomalous difference on the Impersonal Opinion Scale to an acquiescence set. More than likely, it arose from some unknown and uncontrolled differences between the experimental and control groups which were unrelated to the experiment proper. Alternatively, it may have been due to pure chance.

Impact of Computer Knowledge and Experience

One final issue relates to computer knowledge and experience. Analysis of the post-experimental questionnaire indicated that our attempt to achieve an approximate balance between treatment groups on this factor had been unsuccessful. We reviewed answers given to the three post-experimental questions concerning a) self-rated computer knowledge; b) self-rated computer experience; and c) number of semester-length courses related to computer technology. A joint index

was constructed by computing a weighted average of each subject's three numerical answers. Weights were chosen for each question in proportion to the mean score (averaged over all subjects) on that question so as to make all three questions contribute about equally to the joint index. Fifty-nine of the sixty subjects were ranked on this index (one experimental subject neglected to answer any of the questions), and subjects were classified as "high" if they fell above the median and "low" if they fell at or below the median. This classification was then cross-tabulated with the two treatment groups.

In Table 4, we observe a highly uneven distribution of knowledge and experience among the two treatment groups. A relatively large

Insert Table 4 about here.

number of subjects with high knowledge and experience fell into the experimental group, while many of those possessing less knowledge and experience were assigned to the control group. A Chi Square test was performed on these frequencies which resulted in a Chi Square value of 2.04, significant at the .154 level (two-tail). A more sensitive statistical test, the Mann-Whitney test, was performed directly upon the two samples of index values. This test indicated that such an uneven distribution among the experimental and control groups could only have occurred by chance alone with probability .023 (two-tail).

On the basis of these somewhat surprising results, we suspected that the index constructed out of the three post-experimental questions

might have been influenced excessively by the recent experimental experience itself. In other words, we suspected that experimental subjects, fresh from working first-hand with a computer, might rate themselves as excessively knowledgeable and experienced. If true, this would have accounted, at least in part, for the unbalanced frequencies shown in Table 4. However, this did not seem to be the case judging from a cross-tabulation of the number of semester-length computer courses by treatment groups (see Table 5).

Insert Table 5 about here.

An analysis of the frequencies in Table 5 yielded a Chi Square value of 4.32, which is significant at the .038 level (two-tail). A Mann-Whitney test performed directly upon the numerical responses given by the two groups yielded an analogous difference significant at the .047 level (two-tail). The biasing effect of recent experience with the computer would seem far less likely to generate these types of results, since the question asked here was a factual one rather than a request for the subject to rate himself. In light of this, it was concluded that our attempt to create a balanced experimental design had been genuinely unsuccessful.

Fortunately, however, the kind of imbalance created, in conjunction with the previously reported results of the experiment, turned out to be a blessing in disguise. Recall that our original purpose in trying to balance the samples with respect to computer knowledge and experience was to control for the mitigating influence of this

factor upon the predicted effects. Yet the predicted effects were observed, even though the experimental group demonstrated a significantly greater degree of knowledge and experience. This afforded greater rather than less confirmation of our original hypotheses.

SUMMARY AND CONCLUSIONS

An experiment was performed to determine the bias-reducing effect of administering a social science questionnaire by means of a computer. It was assumed that subjects would regard typing answers directly into a computer as a situation guaranteeing them greater privacy and a greater sense of anonymity and confidentiality than the conventional situation wherein questionnaires or psychological tests are filled out with pencil and paper and must then be scrutinized, scored, and interpreted directly by other human beings. Based upon this assumption, two specific hypotheses were formulated. First, whenever the content of a question is regarded by a subject as highly personal and possibly disturbing, he will respond with greater honesty and candor under computer administration as opposed to conventional administration of the questionnaire. Second, when an impersonal and emotionally neutral question is asked, no such difference in response tendency will occur.

To test the above pair of hypotheses, a questionnaire was formulated containing two types of questions. Some of the questions were specifically designed to be highly personal and disturbing. Others

were designed to be impersonal and emotionally neutral. Measures of honesty and candor were constructed out of potential responses to these two types of questions. Specific predictions were then made in terms of these measures concerning differences in the honesty and candor with which subjects would respond depending upon whether the questionnaire was administered by a computer or in the conventional, pencil and paper manner.

A sample of sixty undergraduate engineers was drawn from the senior class of the Massachusetts Institute of Technology. This sample was divided randomly, but evenly, into an experimental group, to receive computer administration of the above questionnaire, and a control group, to receive conventional administration of the same questionnaire.

Results of the experiment confirmed both of the original hypotheses. A small, but statistically significant difference between the manner in which experimental and control groups answered the personal questions did emerge. Furthermore, this difference was in the predicted direction (i.e., the experimental group displayed greater honesty and candor in their responses). In contrast, no systematic difference emerged between the experimental and control groups on the impersonal and emotionally neutral questions.

Additional analyses were performed on three distinct fronts. First, an attempt was made to verify that the experimental situation had in fact succeeded in creating the differential preconditions to which observed differences in response bias had been attributed.

Second, an attempt was made to rule out the possibility that these same results might have been generated by an altogether different phenomenon--the operation of an acquiescence set. Third, elaborations and implications of the original hypotheses were sought. On the basis of these additional analyses, it was concluded that:

- 1) The personal questions included in the questionnaire for the purpose of disturbing subjects and inducing reluctance on their part to answer honestly had in fact been regarded by all subjects as distinctly more disturbing than the impersonal questions designed to be emotionally neutral.
- 2) Experimental subjects indicated that they were slightly more persuaded than control subjects that their answers would be treated anonymously and confidentially.
- 3) There was little evidence to support attribution of observed differences in response bias to the operation of an acquiescence set.
- 4) There was little evidence to support a possible connection between an individual's felt need for privacy and his tendency to bias responses to personal questions.
- 5) Predicted differences in response bias were observed even though the experimental subjects appeared distinctly more knowledgeable of and experienced with computers.

Although the experiment reported herein succeeded in verifying our two hypotheses concerning response bias, the results obtained cannot be regarded as highly definitive. Recall how small most of the reported differences turned out to be. The fact that they achieved a fair degree of statistical significance merely indicates that a phenomenon of the type we hypothesized does exist. However, mere statistical significance does not necessarily indicate that this phenomenon is either large in magnitude or important in its implications. Further research is required to answer both of these more interesting questions.

Concerning the possibilities for further research, at least three are noteworthy. First, even though there was little evidence to support the operation of an acquiescence set in our experiment, a future questionnaire might well be designed in such a way as to provide specific control of this factor in advance. Substitution of multiple-choice questions for true-false questions would be one way to accomplish this end. Randomization of the scoring of any remaining true-false questions would help achieve the same results.

Second, subjects should be used as their own controls in a test-retest experimental design. That is, instead of assigning two separate groups of subjects to the experimental and control treatments, respectively, and then comparing their responses, we would suggest the following procedure:

- 1) Select a sample of subjects.
- 2) Divide the sample randomly, but evenly, into an experimental and control group.

3) After performing the experiment, allow sufficient time to elapse so that subjects will forget the answers they gave on the questionnaire.

4) Repeat the experiment, but with the former experimental subjects now playing the role of controls and vice versa.

This procedure would provide greater control over extraneous between-treatment variations in observed differences and, therefore, permit more sensitive tests of the original hypotheses.

Third, and most important of all, we would suggest transforming the experiment from a single factor to a two-factor design. The first factor would be "the extent to which questions are regarded as personal, disturbing, and emotionally charged." The second factor would be "the extent to which respondents perceive themselves as protected from any ill effects which might result from giving honest and candid answers." The phenomenon to be studied would once again by a form of response bias such as lack of honesty and candor and the role of computers in reducing bias. By restructuring the problem in this manner, it would be possible to observe the individual effects of both factors at various levels, as well as any interaction effects which might exist between the two. In particular, effects of the first factor (degree of personalness) might be investigated at much higher levels of intensity. Subordinates might be asked to evaluate the competence of their supervisors with and without their supervisors' knowledge. This might be particularly effective in generating large differences if supervisor-subordinate pairs were drawn from large, formal organizations

such as the military.

Finally, if subsequent research confirms the findings of this exploratory experiment which, to our knowledge, is the first of its kind, it could have a general implication for social science research. Whenever questions of a personal and salient nature are asked in a sample survey, in a psychological test, or in a laboratory experiment--and this is almost invariably the case in social science research--it would be important to consider the use of computers as a data-gathering instrument. Thus, for example, social scientists engaged in research on formal organizations as well as members of personnel departments conducting employee surveys might profit from using computers for this purpose.

Another area of potential application, if the findings of this experiment are substantiated, is in public opinion surveys. Unlike market research surveys, which are concerned with attitudes and behavior patterns that are not very personal in nature, public opinion surveys often involve delicate and controversial matters. In line with the hypotheses and findings of this study, we would expect more response bias in the case of public opinion surveys than in market research surveys. Hence, the use of computers in administering public opinion surveys--possibly with the help of a special mobile data-phone set--might counteract response bias.

The use of computers in social science research has widely diffused in recent years. Thus far, however, it has been employed almost

exclusively as a data-processing instrument. A new use of computers has been recently explored to provide immediate feedback to respondents of the results of survey research.¹⁷ However, if our findings are confirmed in further experiments, we will have uncovered, in effect, a new function for computers, namely, as a data-gathering instrument for social science research.

Table 1. Scale Results by Treatment

Scale	Favorable* Proportions, Experimental Group	Favorable* Proportions, Control Group	Difference in Proportions	Significance of Difference**
Manifest Anxiety Scale (MMPI)	$\frac{300}{600} = .500$	$\frac{268}{600} = .447$	.053	.019 (1-tail)
Lie Scale (MMPI)	$\frac{67}{450} = .149$	$\frac{81}{450} = .180$	-.031	.075 (1-tail)
Srole Scale, dichotomized	$\frac{21}{150} = .140$	$\frac{18}{150} = .120$	.020	.293 (1-tail)
Factual Scale (Appendix II)	$\frac{263}{300} = .877$	$\frac{274}{300} = .913$	-.036	.114 (2-tail)
Impersonal Opinion Scale (Appendix III)	$\frac{198}{300} = .660$	$\frac{175}{300} = .583$	.077	.011 (2-tail)

* "Favorable" means indicative of the characteristic which the scale purports to measure. For example, a response "favorable" to the Manifest Anxiety Scale would be one indicating manifest anxiety.

** See footnote 14.

Table 2. Pooled Results of Personal vs. Impersonal Questions by Treatment

Favorable* Proportion, Experimental Group	Favorable* Proportion, Control Group	Difference in Proportions	Significance of Difference**
$\frac{704}{1200} = .586$	$\frac{655}{1200} = .545$	.041	.005 (1-tail)
$\frac{461}{600} = .768$	$\frac{449}{600} = .748$	.020	.292 (2-tail)

Pooled results of personal questions with a difference predicted

Pooled results of impersonal questions with no differences predicted

* "Favorable" means indicative of the characteristic which the scale purports to measure. For example, a response "favorable" to the Manifest Anxiety Scale would be one indicating manifest anxiety.

** See footnote 14.

Table 3. Responses on the Factual Scale Where the Factually
Correct Answer Was to Check "False"

	Experimental Group	Control Group
Number of Responses Checked "True"	15	11
Number of Responses Checked "False"	135	139

Table 4. Computer Knowledge and Experience by Treatment

Degree of Knowledge and Experience

	High	Low
Experimental Group	16	13
Control Group	10	20

Table 5. Number of Semester-Length Computer
Courses by Treatment

	None	At Least One
Experimental Group	6	23
Control Group	15	15

APPENDIX I

Privacy Scale

1. I would not like to take a job which required sharing an office with somebody else.
2. I often find it helpful to take a walk by myself so I can think without interruptions.
3. I don't mind telling other people about my salary.
4. I think everybody in the United States should be fingerprinted.
5. When I am in a crowd, I like to lose myself.
6. I enjoy getting telephone calls from survey organizations asking me whether I am watching a particular TV show or listening to a particular radio station.
7. I worry about all the information employers and the government have about people in this country.
8. I don't mind getting unsolicited mail inviting me to join book clubs or buy insurance or things like that.
9. I keep all my correspondence under lock and key.
10. I feel more confident knowing that I am talking on a telephone that has no extensions.
11. I get very upset when someone opens my mail.
12. I don't tell my friends anything about my family.
13. In selecting and promoting executives, I think an employer has the right to consider the personality of each perspective candidate's wife.
14. The Internal Revenue Service should have unrestricted access to any and all information pertaining to business transactions among citizens.
15. Administrative officers in universities should have the right to ask professors for personal information about their students when handling disciplinary matters.

APPENDIX II

Factual Scale

1. There are seven continents on earth.
2. The earth is closer to Mars than to the moon.
3. The number of senators from each state is uniform, but the number of members of the House of Representatives depends upon the population of a state.
4. Doctors are not in short supply in rural areas.
5. The American buffalo is becoming extinct in the Western states.
6. The printing press is over a thousand years old.
7. There were originally 13 states in the United States.
8. Michelangelo was a contemporary of Julius Caesar.
9. The French Revolution preceded the American Revolution.
10. Head hunting is still practiced in some parts of the world.

APPENDIX III

Impersonal Opinion Scale

1. I don't like shopping in supermarkets.
2. I don't like Sunday newspapers that take all day to read.
3. I sometimes find myself puzzled by some works of modern art.
4. I think book reviewers are sometimes too brutal in their criticisms.
5. I think grade schools should teach children at least one foreign language.
6. I think people should be patriotic.
7. I think too many people in America watch baseball games.
8. I think people should be honest in reporting their income to the government.
9. I think more effort should be made to cut down the number of automobile accidents.
10. I think scientific progress will benefit all mankind.

APPENDIX IV

Post-Experimental Questionnaire

- For each of the 120 questions you have just answered, please check your answers to each of the following pair of questions.

<u>Question number</u>	<u>Did you find this question disturbing in any way?</u>	<u>Were you reluctant to give a candid answer to this question?</u>
------------------------	--	---

[120 spaces provided for Yes-No responses]

- How persuaded are you that the answers you have given on the questionnaire will, in fact, be anonymous and confidential?

1	2	3	4	5	6	7
very much			don't			not at all
persuaded			know			persuaded

- How serious do you think you were in expressing your opinions on the items in the questionnaire?

1	2	3	4	5	6	7
very			somewhat			not at all
serious			serious			serious

- How many semester-length courses have you taken which were directly concerned with digital computer technology?

_____ number of courses

- How would you rate your degree of knowledge concerning computers?

1	2	3	4	5	6	7
none			moderate			very
at all						extensive

- How much first-hand experience have you had using computers?

1	2	3	4	5	6	7
none			moderate			very
at all						extensive

FOOTNOTES

* Work reported herein was supported in part by Project MAC, an M.I.T. research program sponsored by the Advanced Research Projects Agency, Department of Defense, under Office of Naval Research Contract Number Nonr-4102(01). Reproduction in whole or in part is permitted for any purpose of the United States Government. The authors are indebted to the Sloan fund of the Sloan School of Management of M.I.T. for a grant which made this study possible. They also wish to express their gratitude to John D. Roach for his valuable assistance in running the experiment and to Jenne Britell for indispensable research assistance.

1. William H. Whyte, Organization Man, New York: Doubleday-Anchor, 1956, pp. 449-456. See also Marvin E. Shaw, "The Effectiveness of Whyte's Rules: 'How to Cheat on Personality Tests,'" Journal of Applied Psychology, 46 (No. 1, 1962), pp. 21-25.
2. Lee J. Cronbach, "Response Sets and Test Validity," Educational and Psychological Measurement, 6 (Winter, 1946), pp. 475-495. See also, William E. Broen, Jr. and Robert D. Wint, "Varieties of Response Sets," Journal of Consulting Psychology, 22 (June, 1958), pp. 237-240.
3. Lee J. Cronbach, "Further Evidence of Response Sets and Test Design," Educational and Psychological Measurement, 10 (Spring, 1950), pp. 3-29.

4. Allen L. Edwards, "Social Desirability and the Description of Others," Journal of Abnormal and Social Psychology, 59 (November, 1959), pp. 434-436; Allen L. Edwards, The Social Desirability Variable in Personality Assessment and Research, New York: The Dryden Press, 1957.
5. Cf. Herbert G. Heneman, Jr. and Dale Yoder, "Employee Opinion Survey by Remote Control," Personnel Journal, 32 (October, 1953), pp. 169-173; Marvin D. Dunnette and Herbert G. Heneman, Jr., "Influence of Scale Administration on Employee Attitude Responses," Journal of Applied Psychology, 40 (April, 1956), pp. 73-77; Leonard I. Pearlin, "The Appeals of Anonymity in Questionnaire Response," Public Opinion Quarterly, 25 (Winter, 1961), pp. 640-647.
6. Martin Orne, "On the Social Psychology of the Psychological Experiment," American Psychologist, 17 (November, 1962), pp. 776-783.
7. A. Vidich and J. Bensman, "The Validity of Field Data," Human Organization, 13 (Spring, 1954), pp. 22-23.
8. Emily L. M. Jones, "The Courtesy Bias in Southeast Asian Surveys," International Social Science Journal, 15 (No. 1, 1963), pp. 70-76.
9. Ibid., pp. 72-73.
10. Cronbach, "Response Sets and Test Validity, op. cit., p. 488. See also, Irwin A. Berg and Gerald M. Rapaport, "Response Bias in an Unstructured Questionnaire," Journal of Psychology, 38 (October, 1954), pp. 476ff.

11. Leo Srole, "Social Integration and Certain Corollaries: An Exploratory Study," American Sociological Review, 21 (December, 1956), pp. 709-716
12. A few additional items were included in the final questionnaire. However, these were not really a part of this experiment, and they were never analyzed.
13. This practice program carried experimental subjects through the same sequence of operations as did the main program, also written in FORTRAN, controlling the question and answer process. Subjects were instructed by the program to type their answers directly into the console, and answers were stored in tape files created by the computer. These tape files were later retrieved for purposes of analysis.
14. In computing statistical significance, the component questions in any given scale were not pooled in order to get a single estimate of the underlying response probability associated with that entire scale. Rather, each component question was treated separately, experimental and control samples were pooled to arrive at a single estimate of the underlying response probability associated with that question, and then results were averaged across questions. This made for a somewhat more **sensitive** test of proportion differences, since it served to control some of the between-question variability.

15. For a related finding on the differential responses to threatening and non-threatening questions, see, Dunnette and Heneman, op. cit., pp. 75-76.

16. We had originally intended to perform a similar analysis on responses to the reluctance question contained in the post-experimental questionnaire. Unfortunately, however, there was substantial evidence that subjects refused to answer this question carefully, if at all. Consequently, no analyses were performed, not even on the very few responses which appeared valid.

Cf, William M. Evan, "Swift Feedback in Survey Research," The Technology Review, 67 (June, 1965), pp. 1-3.

MAY 3 1967

21

NOV 1 1967

APR 1 1968

JAN 5 '69

FEB 27 '69

~~MAR 27 '69~~

SEP 25 '69

~~OCT 9 1969~~

NOV 17 '69

JAN 27 '70

JUN 5 '70

MAR 2 '71

